RESEARCH IN
Comparative &
International Education

Article

Research in Comparative &
International Education
2016, Vol. 11(3) 334–349
© The Author(s) 2016
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1745499916664818
rci.sagepub.com
SAGE

Educational tracking, inequality and performance: New evidence from a differences-in-differences technique

Jeroen Lavrijsen
HIVA–KU Leuven, Belgium

Ides Nicaise
HIVA–KU Leuven, Belgium

Abstract
One of the important differences between educational systems from different countries is the age at which students are placed into separate tracks. We examined the effects of the age at which tracking occurred on student achievement in a comparative perspective, making use of recent waves of three internationally standardized student assessments (PISA, TIMSS, and PIRLS). In order to control for unobserved national heterogeneity, we adopted a differences-in-differences approach, in which we controlled secondary school results for differences already present in primary school (i.e. before the introduction of tracking). The results indicate that early tracking has a negative effect on mean performance of students, particularly in the domain of literacy. Moreover, by separating out groups with different abilities, it is shown that early tracking has a very strong negative effect on low achieving students, suggesting that disadvantageous peer- and environmental effects in the lower tracks may have detrimental consequences on students' academic achievements. By contrast, a null effect on the group of top achieving students was found, suggesting that comprehensive systems can equally challenge high performers to learn at a high pace.

Keywords
Cognitive achievement, diff-in-diff, dispersion, literacy, numeracy, tracking

Studying educational tracking from a comparative perspective

Around the world, educational systems have developed different ways to deal with differences in the capacities of pupils. A common approach in many countries has been to place pupils with different abilities in separate tracks (also known as 'streams'), usually academic or vocational in nature (Field et al., 2007). However, the age at which tracking ('streaming') takes effect differs

Corresponding author:
Jeroen Lavrijsen, HIVA–KU Leuven, Parkstraat 47 bus 5300, B-3000 Leuven, Belgium.
Email: jeroen.lavrijsen@kuleuven.be

considerably: some countries do not track students until age 16, others have different tracks starting at age 10.

Differences in the age at which tracking is introduced between countries are of importance with regard to comparative educational research, because tracking regimes have been shown to influence systematically the functioning of the educational system in a number of ways (Bol and Van de Werfhorst, 2013; Van de Werfhorst, 2014). For example, tracking students at an early point in their career can increase the cognitive efficiency of the educational system, because sorting pupils according to ability facilitates tailored instruction at the right level and pace for every student (Figlio and Page, 2002). In contrast, early tracking might reduce equality of opportunity, because socially disadvantaged students are often disproportionally selected into supposedly less prestigious tracks (Van de Werfhorst and Mijs, 2010), reducing social cohesion in the longer term (Green et al., 2006). In response to some of the assumed negative effects of early tracking, a number of Western countries have, over the last decades, sought to postpone tracking (Antikainen, 2006; Baldi, 2012; Horn, 2007).

Our paper focuses on the effect of early tracking on cognitive performance; it approaches this issue from a comparative perspective. Rooted in the 'methodological empiricist' tradition (Noah and Eckstein, 1969), we first develop a general perspective on the effects of tracking on the basis of the available insights from existing country-specific and trans-national research. We then test this perspective empirically on the basis of trans-national data on student achievement. Thus, rather than comparing and contrasting one particular national educational system with another, our aim is to develop a general account of how the characteristics of the educational system influence its functioning in terms of educational achievement. The objective of this approach is thus, in the words of Harold J Noah (1973), '…not, as in traditional comparative studies, to extend and enrich as far as possible, the connotational content of country-names; instead, we seek to extend and enrich to the limit general law-like, cross-system statements'. In this way empiricist comparative studies, like the present paper, are therefore attempts 'to replace as far as possible the names of systems (countries) by the names of concepts (variables)' – with early tracking being the variable of interest in this paper (Noah, 1973).

The paper is structured as follows. First, we discuss the different mechanisms by which early tracking could influence the educational performance of different groups of pupils. We then summarize the current empirical evidence on the net effects of tracking on educational performance, and then indicate how our paper seeks to contribute to this existing body of knowledge. This is followed by a description of the data and the methodology used; and then the results of our empirical analysis are presented. The paper concludes with a discussion of the results and the implications for future research.

Multiple effects of tracking on cognitive performance

In general, there is some support for the arguments that early tracking can have positive and negative effects. On the positive side, it could be argued that when pupils are sorted according to ability this makes instruction at the right level and pace more feasible. This would boost the efficiency of the educational system and increase the average performance (Hallinan, 1994a, 1994b; Duflo et al., 2008). This 'specialisation benefit' applies to both high and low achieving students, and hence should increase performance over the full ability distribution.

However, this view has been repeatedly challenged: in particular, it is argued that weak students could suffer from early tracking, because shifting these pupils to less demanding tracks would lead to their problems being ignored rather than being addressed (Hallinan and Kubitschek, 1999; Hattie, 2002). The mechanisms suggesting such an adverse effect of tracking on low performing students have been summarized as follows by Oakes (1993): ‘Students not in the highest tracks have fewer intellectual challenges, less engaging and supportive classrooms, and fewer well-trained teachers’. This perspective stresses the influence of the ability level of classroom peers on individual performance. However, the size and direction of these peer effects is not yet fully clear: opponents of early tracking maintain that weak students benefit from the presence of stronger peers (Hanushek et al., 2003; Hoxby, 2000), while proponents argue that low-ability students are better off in the company of peers of their own level (Dobbelsteen et al., 2002). It has also been suggested that tracking gives rise to the development of class and school cultures that are negatively oriented towards learning in the lowest tracks: students in these tracks often end up there because of negative selection, and this experience of failure may induce a ‘sense of futility’ in the classroom that is detrimental to the learning climate (Van Houtte and Stevens, 2009, 2010).

In addition, it has been suggested that educational resources are unequally distributed across tracks (Darling-Hammond, 1995). For example, it has been repeatedly shown that, particularly in differentiated systems, the more experienced and more capable teachers are often assigned to higher tracks in advantaged schools (see OECD (2012) for recent trans-national data). Moreover, teachers working with lower track individuals may develop lower expectations towards their students and act likewise (Van Houtte, 2004), e.g. by devoting less time to actual instruction in lower tracks (Hallinan, 1994a).

While such mechanisms suggest negative effects of tracking in the lower tracks, they could work in the opposite direction in the upper tracks: high performers would not only benefit from the focussed instruction in homogeneous groups, but also enjoy having peers, teachers and other educational resources of the highest quality. However, it is to be noted that tracking is not the only way to deal with student heterogeneity. For example, the Nordic comprehensive systems have adopted successful mechanisms, such as within-class differentiation and high quality remedial education, to differentiate instruction by ability level without having to revert to rigid tracking (Dupriez et al., 2008). Such mechanisms allow schools and teachers to maintain a high pace of instruction, without losing the low performers.

Finally, it has been argued that the assignment of young pupils to tracks is far from noiseless (Dustmann, 2004). At a young age, social background often biases track placement: see, for example, Boone and Van Houtte (2012) for Flanders, Ditton and Krusken (2006) for Germany, and Duru-Bellat (2002) for France. Misallocations of students to the ‘wrong’ track might mean that tracks are less homogeneous (in terms of ability) than is assumed by the specialisation benefit; furthermore, it would imply that not all talented but socially disadvantaged pupils would get the opportunity to fulfil their full potential.

Existing empirical evidence on the net effect of tracking

The mechanisms discussed above exert conflicting pressures: some might affect the cognitive performance of a particular group of students in a positive way, while other mechanisms suggest negative effects. The net effect of tracking will therefore depend on the relative strength of the different mechanisms at work. In what follows we will review some studies that have attempted to determine empirically the net effect of tracking on performance.

Single-country studies

First, empirical research in a single country exploits within-country differences in tracking regimes to estimate the effect of tracking on performance. For example, a series of studies has considered local variations in policies regarding tracking in high schools in the USA. In general, this research suggested that tracking was slightly beneficial for high achievers, but at the expense of lower performance for low achievers (Argys et al., 1996; Hallinan and Kubitschek, 1999; Hoffer, 1992). However, other researchers have found achievement benefits for all students, even those of lower ability (Figlio and Page, 2002; Galindo-Rueda and Vignoles, 2004).

A drawback of such within-country research is that it has proved difficult to control out all *other* differences between tracked and non-tracked schools. Indeed, tracked and non-tracked schools often tend to differ in other aspects as well: for example, they cater for different socio-economic school populations, or they differ in the kind of teachers they recruit. Unmeasured factors correlated to differences in tracking regimes could bias the estimate of the effect of tracking. For instance, it has been shown (Manning and Pischke, 2006) that the positive effect of tracking on performance observed by Galindo-Rueda and Vignoles (2004) was driven mainly by selection bias, because the implementation of tracking was correlated to previous cognitive achievement.

Comparative cross-sectional studies

A solution to the problem of selection bias inherent to within-country studies has been to exploit cross-national variation in tracking regimes, i.e. by comparing representative samples of the full student population from both early and late tracking countries. For example, in Germany students are tracked after Grade 4 in three school types (Hauptschüle, Realschüle, Gymnasium), while in Nordic Europe comprehensive (non-tracking) schooling lasts until age 16. Thus the question arises: how does the educational achievement of German students compare with that of their Nordic counterparts? Comparative research of this type has been strongly facilitated by the proliferation of several international large-scale student assessments during the last decade, which deliver the standardized data needed to compare achievement across countries adequately.

However, trans-national studies of this kind have to deal with their own important pitfall. A central proposition of classical comparative research is that educational systems do not exist independent of their social or economic context but, rather, are strongly related to it. As Michael Sadler (1964) put it: ‘In studying foreign systems of education we should not forget that the things outside the schools matter even more than the things inside the schools, and govern and interpret the things inside’. Hence, in comparing educational systems from different countries, we should address with some caution the many differences, other than the variable of interest, that might also influence educational performance. Statistically, this corresponds to attempting to control out the bias induced by ‘confounding variables’ in order to provide unbiased estimates.

The best studies in this tradition have therefore tried to accommodate for the possible heterogeneity bias by including a list of possible confounding variables in their model. In general, two groups of candidate variables can be identified: first, there are socio-economic and cultural factors which may influence educational outcomes; and, second, a set of features of the educational system itself (e.g. school autonomy) has been shown to influence performance as well. For example, Horn (2009) examined the effect of tracking age on average student performance in PISA 2003 with a multilevel set-up. However, while other features of the educational system were kept under control (e.g. school autonomy, accountability, size of vocational education), the possible bias induced by differences in the socio-economic context of nations was not addressed. By contrast, while Duru-Bellat and Suchaut (2005) analysed PISA 2003 with GDP and the educational level of society as context controls, the bias possibly induced by other characteristics of the educational system was neglected.

Overall, these and other studies (Dupriez et al., 2008; Schütz et al., 2008; Van de Werfhorst and Mijs, 2010; Woessmann et al., 2009) reported negative or null effects of early tracking on average performance. However, one study (Rindermann, 2007) came to an opposite conclusion. By aggregating national scores over an array of student assessments from different sources, Rindermann (2007) found a positive effect of early tracking on the average score; but the aggregation led to the construction of samples outside the typical scope of Western countries, and a reanalysis of the data (Lavrijsen and Nicaise, 2009) showed that the reported positive effect vanished when the analysis was restricted to OECD-countries. This example thus shows that adequate context control is ultimately critical with regard to the reliability of trans-national studies.

The problem with any attempt to remove the bias by country-level confounders, however, is that one can never guarantee that *all* confounders have in fact been adequately taken into account. Moreover, some of the supposed confounders may be very difficult to measure reliably (e.g. the cultural value attributed to education). Equally, confounding variables may have non-linear effects; for example, it has been argued that the size of the educational budget does not influence performance once a certain threshold has been exceeded (Hanushek and Luque, 2003). A final restriction is that because of limited sample size (with typical samples of between 20–30 countries) a simultaneous control for several confounders in one model is unfeasible. Hence, when the reported studies seemed to suggest a negative or null effect of early tracking, uncertainty and caution about the validity of this finding remains in place.

Differences-in-differences techniques

A promising strategy to deal with the issue of unobserved national heterogeneity is the differences-in-differences approach (‘diff-in-diff’). Essentially, diff-in-diff corrects differences between countries in secondary school outcomes for differences already existing in primary school. The results in primary school can be assumed to be influenced by the same unobserved variables as the results in secondary school (culture, the social context, etc.). Under this assumption, diff-in-diff would thus remove all bias introduced by confounding variables, without having to include the cofounders in the model themselves.

A frequently cited example exploiting such a diff-in-diff approach to estimate the effect of early tracking on achievement is presented in the paper by Hanushek and Woessmann (2006), who compared the academic performance of students from different countries at age 15 with the performance scores of students in primary school (age 10). Because only the measurement point at age 15 is influenced by tracking age (no countries track students before age 10), the net effect of tracking on performance can be determined on the basis of the difference between early and late tracking countries in the increase in academic performance between both measurement points. Drawing on eight combinations of a primary with a secondary school assessment (see Table 1), Hanushek and Woessmann (2006) estimated the effect of tracking on the performance of the average student, the performance of a group of low achievers, the performance of a group of high achievers and, finally, the differential between low and high achievers. In the discussion of their results, however, Hanushek and Woessmann (2006) mainly focussed on the latter outcome. Indeed, their data showed convincingly that early tracking amplified the differential between weak and strong students; in all of their eight specifications, tracking led to larger gaps between low and high performers. However, the results regarding mean performance were more mixed: three combinations indicated a sizeable negative effect of early tracking, but four other combinations produced a null effect, and one combination even yielded a significantly *positive* estimate (Table 1, last column). Thus while Hanushek and Woessmann (2006) concluded that there is a ‘tendency’ for early tracking to reduce mean performance, they admitted that this part of the conclusion was ‘less clear’.

As discussed above, the central issue for the opponents of early tracking is that tracking would harm the (absolute) performance of low achieving students – those placed in the less prestigious tracks. However, even in this regard the results from Hanushek and Woessmann are not

Table 1. The eight combinations of primary and secondary assessments studied by Hanushek and Woessmann (2006).

Model	Primary school assessment	Secondary school assessment	Domain	*N*	Estimate of the average effect of tracking
A	PIRLS 2001 (4th grade)	PISA 2003 (15-year-olds)	Reading	18	−1.1***
B	PIRLS 2001 (4th grade)	PISA 2000 (15-year-olds)	Reading	20	−1.0***
C	TIMSS 1995 (4th grade)	TIMSS 1995 (8th grade)	Math	26	–0.1
D	TIMSS 1995 (4th grade)	TIMSS 1995 (8th grade)	Science	26	0.6**
E	TIMSS 2003 (4th grade)	TIMSS 2003 (8th grade)	Math	25	–0.0
F	TIMSS 2003 (4th grade)	TIMSS 2003 (8th grade)	Science	25	–0.0
G	TIMSS 1995 (4th grade)	TIMSS 1999 (8th grade)	Math	18	–0.4*
H	TIMSS 1995 (4th grade)	TIMSS 1999 (8th grade)	Science	18	0.2

*p < 0.1.
**p < 0.05.
***p < 0.01.

unambiguous: they found that in most aspects weak students in early tracking countries did perform markedly worse than their counterparts in late tracking countries (Hanushek and Woessmann, 2006). However, in one aspect low performing students again seemed to *benefit* from early tracking. Even when this benefit was still smaller than the benefit of tracking for high performers (again leading to a larger gap between weak and strong students), this finding seems at odds with expectations.

Hanushek and Woessmann (2006) interpreted the amplifying effect of tracking on the gaps between weak and strong students as an indication that non-tracking of schools would result in better outcomes, in particular on the lower end of the ability distribution. However, if the tendency observed in this single aspect should prove valid, it is hypothesised that this would alter the policy implications of their research: when *every* student would gain (in absolute terms) from early tracking, increased gaps between weak and strong performers (in relative terms) might become acceptable, because the alternative of non-tracking would cause everyone to lose. However, Hanushek and Woessmann (2006) gave little attention to this inconsistency, arguing that, ultimately, in all combinations weak students were worse off under early tracking compared to stronger students.

Our contribution

Our article thus has two principle aims in contributing to the existing literature. First, as discussed above, the bulk of the existing empirical trans-national research has exploited cross-sectional designs, but the essential problem of such studies – the possible bias introduced by unobserved national-level confounding factors – can never be completely discarded. As we have argued above, we propose a diff-in-diff-approach as the most effective tool to deal with this. Second, although the work by Hanushek and Woessmann (2006) had adopted the diff-in-diff-approach, some problems remain to be solved. In particular, we will focus more on the effect of tracking on the absolute performance of both low and high achievers, rather than on the gap between such individuals.

There are several reasons why we believe that this article will improve on the analysis by Hanushek and Woessmann (2006). First, while the student assessments examined by Hanushek and Woessmann were all conducted between 1995 and 2003, we can now make use of the wealth of data from student assessments that has become available more recently (2001 to 2011). This is important not only because of the timeliness of the analysis but also because it allows us to use larger sample sizes and to increase statistical reliability. Thus while the assessments studied by Hanushek and Woessmann yielded samples sizes of between 18 and 26 countries, the steady increase in the number of countries participating in international student assessments now allows us to construct samples with up to 35 countries.

Second, these new series of student assessments allow us to examine more combinations that have PISA as the secondary school measurement point. Hanushek and Woessmann (2006) reported only two specifications with PISA as the secondary measurement point, with relatively small samples sizes (18–20 countries), while they used TIMSS for the other six specifications (Table 1). However, in a diff-in-diff-setting PISA seems preferable to TIMSS, because the average age of the respondents in PISA (15.8 years) is significantly higher than in TIMSS (14.3 years). Hence, when PISA is used as the endpoint, tracking (which is usually introduced between ages 10 and 12) has had more time to exert its influence and its effects should be more accurately detectable.

Finally, the average age of participants in the different student assessments – and hence the length of the time between the first and the second measurement point, or the length of the performance growth – differs across nations. As noted by Jakubowski (2010), Hanushek and Woessmann (2006) failed to take these age differences between countries into account, which may have distorted their estimates. We will respond to this valid criticism by including the average age of participants for each country in the model.

Data and methodology

All our differences in differences models rely on a relationship of the form:

$$Y_{\text{secondary},i} = a + b.Y_{\text{primary},i} + c.T_i + d.X_i + \varepsilon_i \qquad (1)$$

Here, *Y* is the outcome under study in a primary versus secondary school assessment (e.g. mathematics performance) for country *i, T* is the tracking indicator, and is a normally distributed error term. For each country we include the average difference in test age in *X*, to prevent differences between nations from distorting our estimates. The assumption that other country-level confounding factors do not bias our estimated effect of tracking can be checked by additionally including other cross-country differences in *X*. We are interested in the effect of tracking on both the mean performance and the performance level of different groups in the ability distribution; hence, we run models with the achievement scores at different quantiles in *Y*.

We use data from three different international student assessments series, with results from several waves of each assessment. Our measures for achievement in primary school are obtained from PIRLS (reading – waves 2001, 2006 and 2011) and TIMSS–4th grade (science and mathematics – waves 2007 and 2011), while those for achievement in secondary school are distracted from TIMSS–8th grade (science and mathematics – waves 2007 and 2011) and PISA (reading, science and mathematics – waves 2006 and 2009). Naturally, only performance scores on the same domain (reading, mathematics, or science) can be compared. In total, this yields 26 combinations of primary and secondary measurement points to be examined. However, because only results for countries participating in both measurement points can be used, some of the combinations represent only small datasets. As a result, we focused our attention on the eight combinations with the largest *N*, because these can be expected to provide the most reliable estimates. While we will limit our further discussion primarily to these eight combinations, results for the other combinations were in line with our findings, although – as expected – with lower power (results available on request from the authors).The eight principal combinations are presented in Table 2.

Table 2. Our eight combinations of primary and secondary assessments.

Model	Primary school assessment	Secondary school assessment	Domain	*N*
1	TIMSS 2007 (4th grade)	PISA 2006 (15-year-olds)	Math	23
2	TIMSS 2007 (4th grade)	PISA 2006 (15-year-olds)	Science	23
3	PIRLS 2001 (4th grade)	PISA 2006 (15-year-olds)	Reading	23
4	PIRLS 2006 (4th grade)	PISA 2006 (15-year-olds)	Reading	27
5	PIRLS 2006 (4th grade)	PISA 2009 (15-year-olds)	Reading	30
6	TIMSS 2011 (4th grade)	PISA 2009 (15-year-olds)	Math	35
7	TIMSS 2011 (4th grade)	PISA 2009 (15-year-olds)	Science	35
8	PIRLS 2011 (4th grade)	PISA 2009 (15-year-olds)	Reading	35

As can be seen from Table 2, these combinations all have PISA as the second endpoint. As argued above, using PISA is preferable to its alternative (TIMSS), as the average age of respondents in PISA is higher and hence tracking has had more time to exert its influence on performance. Note that in the study Hanushek and Woessmann (2006) exactly the same two combinations using PISA as the endpoint unfortunately relied on small sample sizes (18 versus 20 countries).

It should be noted that the three different student assessments measure somewhat different types of achievement (Micklewright and Schnepf, 2007). PISA assesses primarily if students are able to use their reading, mathematics and science skills in real-life situations. With regard to reading skills, this is similar to the approach adopted in PIRLS, which is based on a similar expanded notion of reading literacy and measures a comparable type of skills as in PISA (Mullis et al., 2006). In contrast, however, TIMSS focuses more on measuring the extent to which the respondent masters an internationally agreed mathematics versus science curriculum. This is a somewhat different perspective then PISA's 'real life skills'. While we note that this difference does not prevent country performance in PISA and TIMSS from being correlated significantly (Rindermann, 2007), it could be argued that those specifications comparing PISA with PIRLS (i.e. combinations 3, 4, 5 and 8) prove the most informative.

Two other issues regarding data comparability are that, first, the composition of school cohorts may change over time (e.g. through immigration) and, second, the environment in which educational system operates may also change suddenly (e.g. as a result of economic crises). To address the first issue, we include a number of combinations in which the primary and the secondary school assessment both tested the same age cohort, i.e. in which the PISA-assessment was carried out five years after the primary school assessment (e.g. combination 3). The second issue is addressed by including combinations of assessments that were delivered at the same time-point (e.g. combination 4).

Finally, to perform a further check of the comparability of both the primary and the secondary assessment tests, which were not on the same scale, we standardized assessments scores to have mean 0 and variance 1 across the countries participating in both tests. This is similar to the procedure used by Hanushek and Woessmann (2006). However, we will report the results with both the standardized and the non-standardized scores, because the latter may have the advantage of being more readily interpretable, because they can be expressed as gains/losses in terms of points in the PISA-assessment.

The eight principal combinations yield samples of 23–35 countries (see Appendix 1). Note that these samples sizes are considerably larger than the ones used by Hanushek and Woessmann (2006). For each country, we also added tracking ages as reported by the OECD (2009). We distinguished two groups of countries: 'early tracking' countries which track their students in between both measurement points (i.e. before age 15) and countries that do not (models using raw tracking ages instead of the tracking variable did not change our results).

Results

The effect of tracking on mean performance

With specification (1), we can now examine the effect of tracking on the mean performance in our set of countries: Tables 3 and 4 list the results.

As expected, average results in primary school assessments correlate strongly with later secondary school assessments. The age difference between the two tests also features significantly in most models, which validates the caution issued by Jakubowski (2010) on the original results by Hanushek and Woessmann (2006).

For our variable of interest, early tracking, the estimates point to a negative effect of early tracking on average performance. In contrast with the mixed results from Hanushek and Woessmann (2006), our results are more consistent: six out of eight specifications yielded a negative effect of early tracking on mean performance, while the two other produced a null effect. The negative effect becomes significant in two of the specifications, with a maximal difference of 22 PISA-points between early and late tracking countries (Figlio and Page, 2002). This is a sizeable difference, because the effect of one year of schooling is estimated to be around 40 points in PISA (OECD, 2009). We can also merge the information available from the different specifications into one general figure by calculating the weighted average over all specifications, taking into account the number of countries used in each combination. This weighted average indicates a loss in mean performance due to early tracking equal to 9 points. When we take into account in addition the other 18 possible combinations (i.e. those with smaller N), the weighted estimate of the effect of

Table 3. Regression of mean performance on early tracking.

	Model 1 Mathematics	**Model 2 Science**	**Model 3 Reading**	**Model 4 Reading**
Constant	17.58	2.73	−193.54	–206.98
Primary assessment	0.74***	0.71***	0.84***	0.97***
Early tracking	–0.11	–7.65	−10.55	–21.81*
Age difference	17.93	24.48*	42.30**	33.03**
R^2	0.86	0.85	0.61	0.72
N	23	23	23	27
	Model 5 Reading	**Model 6 Mathematics**	**Model 7 Science**	**Model 8 Reading**
Constant	–87.65	–89.21	−159.25	–224.10
Primary assessment	0.75***	0.84***	0.89***	1.01***
Early tracking	−16.17**	0.10	–7.72	–11.99
Age difference	33.56***	25.87**	35.69***	31.49**
R^2	0.81	0.81	0.75	0.76
N	30	35	35	35

*$p < 0.1$.
**$p < 0.05$.
***$p < 0.01.1$.

Table 4. Coefficients of the effect of early tracking on mean performance, after normalisation of scores.

	Model 1	Model 2	Model 3	Model 4
Early tracking	–0.01	–0.15	–0.25	–0.47*
	Model 5	**Model 6**	**Model 7**	**Model 8**
Early tracking	–0.45**	0.00	–0.17	–0.28

early tracking is identical (minus 9 points). Hence, our first finding is that early tracking does not seem to offer benefits for average performance; rather to the contrary. Note that using standardized scores (Table 4) did not alter our results.

Tables 3 and 4 may suggest a possible difference in the effect of early tracking on different subject domains. Indeed, the two combinations referring to mathematical assessments (specifications 1 and 6) yielded only null effects, in sharp contrast with the large negative effects that we observe in the reading assessments (specifications 3, 4, 5 and 8). We cannot determine from the data whether this is genuinely a subject-related issue (e.g. when teaching of mathematics benefits more from homogeneous classrooms then do classroom practices related to reading literacy) or whether this is due to lower comparability of PISA and TIMSS as opposed to PISA and PIRLS (see the methodology section). However, note that when we take into account the other eight combinations regarding mathematical performance (with smaller *N*), the weighted average effect in mathematical performance is minus 7 points. Hence, we do seem to observe a slightly negative effect of early tracking on mean mathematical performance as well, but it seems smaller in size than the effect on literacy.

Finally, we performed a check on one of the central assumptions behind the diff-in-diff approach: whether it really removes the bias of confounding variables such as wealth. For this, we included a control for wealth (GDP/capita) in specification 1. In most cross-national research, the level of a country's economic development has been picked as one of the most influential confounders of cross-country comparisons of educational performance (Duru-Bellat and Suchaut, 2005). However, if wealth influences both primary and secondary school assessments to a reasonably comparable degree, it should not affect the diff-in-diff estimates. This is indeed exactly what we observe: wealth did not prove not to be consistently related (either positively or negatively) with the performance gain between the two measurement points, and the additional control did alter neither the sign nor the size of the effect of early tracking. Weighted over the eight combinations, the effect of early tracking controlled for wealth is estimated to be minus 8 PISA-points, virtually equal to the effect reported above. Note that this of course does not imply that wealth would not exercise a significant effect on assessment scores: it only means that wealth influences both primary and secondary assessments in the same way and thus is not associated with the performance gain. Figure 1 visualizes this phenomenon (for specification 6): wealth is positively associated both with the primary assessment scores and secondary assessment scores, but it does not have a clear-cut association with the *difference* between both (when added to the model, the estimate for wealth had a *p*-value of 0.97). Hence, the diff-in-diff-design adequately precludes wealth to bias the estimation of the effect of early tracking, which seems to confirm its advantage over trying to include explicitly an (unknown) set of possible confounders in the model.

The effect of tracking on weak and strong students

We now turn to the effect of early tracking on different groups in the achievement distribution. We follow the same estimation strategy as above, but our inputs (Y_{primary}) and outcomes ($Y_{\text{secondary}}$) now refer to quantile scores instead of mean scores. This gives an indication of the effect of tracking on

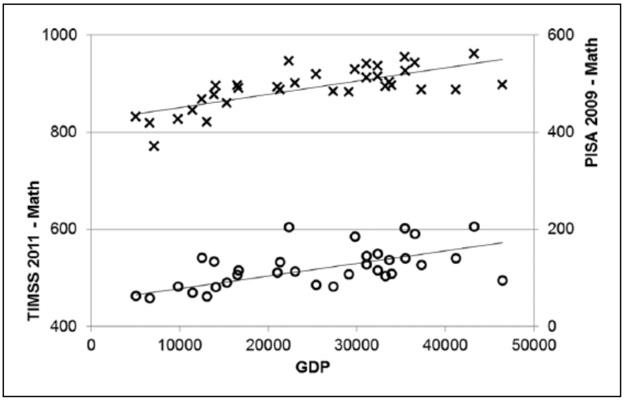

Figure 1. Mathematical performance in PISA2009 (above) and TIMMS–4th grade 2011 (below) as a function of GDP/capita.

Table 5. Coefficients of the effect of early tracking on the scores for different quantiles.

	Model 1 Mathematics	Model 2 Science	Model 3 Reading	Model 4 Reading
Q05	–3.44	–2.24	–9.08	–36.51***
Q25	–3.95	–8.67	–11.75	–26.17**
Mean	–0.11	–7.65	–10.55	–21.81*
Q75	3.63	–8.03	–9.33	–13.79
Q95	9.78	–2.59	–7.52	–9.88
	Model 5 Reading	**Model 6 Mathematics**	**Model 7 Science**	**Model 8 Reading**
Q05	–23.83**	–5.01	–6.71**	–20.51
Q25	–21.82***	–6.51	–11.11**	–19.69
Mean	–16.17**	0.10	–7.72	–11.99
Q75	–10.55	6.02	–3.35	–4.75
Q95	–8.80	15.11	1.30	–0.21

*p < 0.1.
**p < 0.05.
***p < 0.01.

the performance of various groups across the performance distribution. For example, the 25%-quantile (Q25) is the score of the first pupil who scores better than the bottom 25% of the participants. Hence, the effect of early tracking on Q25 indicates the effect of tracking on the performance of the lowest quartile in the achievement distribution.

Table 5 presents the estimates of the effect of early tracking on four separate quantiles in the achievement distribution, ranging from Q5 (very low achievers) to Q95 (very high achievers). First, note that in all specifications, the effects of early tracking are the most negative in the groups with the lowest achievement. This is what we would expect on the basis of peer- and environmental effects, which have a negative effect in particular on those grouped in the lower tracks.

In terms of absolute performance levels, we observe a strong negative effect of early tracking on the performance of lower achievers. The estimates for the 5%- and the 25%-quantile are consistently negative in all specifications and gain statistical significance in three of the combinations. Note the strong consistency of our estimates compared to the results by Hanushek and Woessmann (2006), who reported a performance *gain* for low achievers in one specification. The weighted average of the effect of tracking is minus 14 points in both quantiles. Again, the strongest effects are found for reading literacy, which probably benefits from the high degree of comparability between the types of abilities tested, with an effect up to 36 points, equalling one year of schooling, in specification 4.

Hence, in terms of the results from low achievers, we argue that there should be no doubt about the negative effects of early tracking. Our findings strongly confirm earlier insights from the literature in which caution has been repeatedly expressed regarding the effect of tracking on low-achieving students. It appears that the instructional advantages of homogeneous groupings seem to be outweighed by the mechanisms, such as lower quality peers or the development of negative class cultures, that cause early tracking to be harmful for lower achievers.

In contrast, our estimates for the effect of tracking on high achievers are more mixed. Overall, early tracking does not seem to offer major benefits, even for this group. The effect of early tracking on the upper quarter (Q75) is negative in most specifications, where the weighted effect over all specifications equals minus 5 points. However, none of the specifications yield significant estimates either in the positive or the negative sense.

The same seems to be the case for the effect on the Q95, which is an indication of the level of the top achievers. Three specifications show a positive (but not significant) effect of early tracking on the strongest performers, others still show small negative effects, and the weighted average is effectively zero (+0.1). Again, it is to be noted that the mathematics specifications seem to produce the largest positive effects, but also that we cannot be sure whether this indicates a real benefit or is an artefact of lower comparability between TIMSS and PISA.

In general therefore our results indicate that early tracking is not beneficial even for the strong performers. Note that most of the mechanisms proposed to explain the negative effects of tracking for low performers (environmental and peer effects) should be interpreted as working to the advantage of strong performers. Together with more effective teaching in homogeneous groupings, this should in fact generate tangible advantages for strong performers in tracked systems. That we do not observe such an effect suggests that educational systems can develop more flexible differentiation strategies that challenge high performers just as well as placing them in a rigid track (Dupriez et al., 2008).

Conclusions

We have examined the effects of early tracking on achievement by applying a diff-in-diff design to data from an array of recent waves of student assessments (PISA, TIMSS, PIRLS). Our results showed consistent negative effects of early tracking on mean performance. Thus we conclude that early tracking does not appear to offer specialization benefits to the average student. Moreover, by considering separately the effects on groups with different abilities, we have demonstrated that early tracking had particularly strong negative effects on the group of low achievers. This can be explained by the negative peer- and environmental effects which reduce learning opportunities in the lower tracks. In contrast, we did not find a consistent effect of early tracking on the achievement of top performers. Thus we conclude that comprehensive (non-tracked) systems seem to be able to challenge high performers to learn at a fast pace, without having to isolate such individuals in separate tracks.

However, there are a number of caveats – and possibilities for future research – we would like to raise. First, we noted that the size of the (negative) effect of early tracking seemed to differ depending on the subject domain: it appeared to be smallest in the numeracy assessments and largest in the literacy assessments. Exploration of the extent to which this finding is related to the particularities of the subject itself would be an interesting topic for future research. One explanation of our finding could be that, for a weak student in an highly abstract subject such as mathematics, the benefits of being taught in a homogeneous classroom, in which the teacher can adjust their pace better according to the ability of the student, are relatively more important; while for a student with low literacy skills, their performance will benefit mainly from interaction and communication with more able classroom peers. However, the available data do not allow us to test this idea, because an alternative explanation – that the differential effect is mainly due to data issues, in particular the lower level of comparability between PISA and TIMSS as opposed to that between PISA and PIRLS – also cannot be excluded.

Second, it is to be noted that a large-scale international comparison such as ours must rely on a relatively crude categorization of nation-specific educational practices. We defined tracking in terms of the age of first selection, but national practices are often more subtle than such a quantification might suggest (cf. the multiple forms of 'equivalence' that can be established between practices from different countries as discussed by Stefan Novak (Novak, 1977). For tracking, our particular variable of interest, it has already been demonstrated that some late-tracking countries have 'solved' the problem of heterogeneous classrooms mostly by means of an increased use of grade retention – which has effects that are largely equivalent to those of tracking (Dupriez et al., 2008). Similarly, even within an early tracking system the adverse effects of tracking could be mitigated by making tracking less rigid; for example, in the Netherlands promotion to a 'stronger' track at a later stage in the educational career is facilitated, which obviously reduces the impact of the earlier track placement (Van de Werfhorst, 2014).

This returns us to the central concern in comparative research. The aim of this study was to select one variable with regard to which educational systems differ and attempt to isolate its effect on performance. The question is to what extent it is really possible to 'replace the name of countries by the names of variables' in this sense; the main insight that the design of the educational system cannot be seen as fully independent from its broader socio-economic or cultural context (Sadler, 1964) makes one rather cautious with regard to interpretation of our results. As already proposed by Merritt and Coombs (1977), 'no social science theory has even approached the goal of accounting for all of the variance in the dependent concept, and we shall have to be content with partial theories'. Nevertheless, we hope that our methodological improvement, i.e. the adoption of a diff-in-diff-technique to isolate, as much as possible, the effect of tracking in secondary education, will add to the understanding of the functioning of educational systems in a comparative perspective. In this sense, we hope that our findings, together with a series of earlier results drawn from a range of methodological and disciplinary perspectives (Van de Werfhorst and Mijs, 2010), will contribute further to the establishment of such a 'partial theory' on how tracking in secondary school affects the cognitive performance of weak and strong students.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by the Flemish Government (Policy Research Centre 'Educational and School Careers').

References

Field S, Kuczera M and Beatriz P (2007) No more failures: Ten steps to equity in education. Paris: OECD. Available at: https://www.oecd.org/education/school/45179151.pdf (accessed 01 August 2016).

Figlio DN and Page ME (2002) School choice and the distributional effects of ability tracking: Does separation increase inequality? *Journal of Urban Economics* 51(3): 497–514.

Galindo-Rueda F and Vignoles A (2004) The heterogeneous effect of selection in secondary schools: Understanding the changing role of ability. *IZA Discussion Paper Series*. Bonn: IZA (Forschungsinstitut zur Zukunft der Arbeit GmbH).

Green A, Preston J and Janmaat G (2006) *Education, Equality and Social Cohesion: A Comparative Analysis*. Basingstoke: Palgrave.

Hallinan MT (1994a) School differences in tracking effects on achievement. *Social Forces* 72(3): 799–820.

Hallinan MT (1994b) Tracking: From theory to practice. *Sociology of Education* 67(2): 79–84.

Hallinan MT and Kubitschek WN (1999) Curriculum differentiation and high school achievement. *Social Psychology of Education* 3(1–2): 41–62.

Hanushek EA and Luque JA (2003) Efficiency and equity in schools around the world. *Economics of Education Review* 22(5): 481–502.

Hanushek EA and Woessmann L (2006) Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *Economic Journal* 116(510): C63–C76.

Hanushek EA, Kain JF, Markman JM and Rivkin SG (2003) Does peer ability affect student achievement? *Journal of Applied Econometrics* 18(5): 527–544.

Hattie JA (2002) Classroom composition and peer effects. *International Journal of Educational Research* 37(5): 449–481.

Hoffer TB (1992) Middle school ability grouping and student achievement in science and mathematics. *Educational Evaluation and Policy Analysis* 14(3): 205–227.

Horn D (2007) Conservative states, stratified education, unequal opportunity: A hypothesis on how educational regimes differ. ASPAnet Conference, Vienna, Austria, 26 – 29 June. Washington, DC: ASPA.

Horn D (2009) Age of selection counts: A cross-country analysis of educational institutions. *Educational Research and Evaluation* 15(4): 343–366.

Hoxby C (2000) *Peer effects in the classroom: Learning from gender and race variation*. Cambridge, MA: National Bureau of Economic Research.

Jakubowski M (2010) Institutional tracking and schievement growth: Exploring differencein-sifferences approach to PIRLS, TIMSS and PISA sata. In: Dronkers J (ed.) *Quality and Inequality of Education: Cross-National Perspectives*. Heidelberg and New York: Springer, pp. 41–81.

Lavrijsen J and Nicaise I (2014) Comprehensief onderwijs: een bedreiging voor kwaliteit? Een heranalyse van Rindermann en Ceci (2009). *Pedagogische Studiën* 91(4): 270–279.

Manning A and Pischke JS (2006) Comprehensive versus selective schooling in England and Wales: What do we know? Available at: http://econ.lse.ac.uk/staff/spischke/grammars.pdf (accessed 01 August 2016).

Merritt RL and Coombs FS (1977) Politics and educational reform. *Comparative Education Review*21(2/3): 247–273.

Micklewright J and Schnepf SV (2007) *Inequality of Learning in Industrialised Countries*. Oxford: Oxford University Press.

Mullis IVS, Kennedy AM, Martin MO and Sainsbury M (2006) *PIRLS 2006 Assessment Framework and Specifications, 2nd Edition*. Available at: http://timss.bc.edu/PDF/P06Framework.pdf (accessed 01 August 2016).

Noah HJ (1973) Defining comparative education: Conceptions. In: Edwards R, Holmes B and Van de Graff J (eds) *Relevant Methods in Comparative Education*. Hamburg: UNESCO, pp. 109–117.

Noah HJ and Eckstein MA (1969) *Toward a Science of Comparative Education*. New York: Macmillan.

Novak S (1977) The strategy of cross-national survey research for the development of social theory. In: Szlai A and Petrella R (eds) *Cross-national Comparative Survey Research: Theory and Practice*. Oxford: Pergamon Press.

Oakes J (1993) Creating middle schools: Technical, normative and political considerations. *Elementary School Journal* 93(5): 461–480.

OECD (2010) *PISA 2009 Results: What Makes a School Successful? Resources, Policies and Practices (Volume IV)*. Paris: OECD.

OECD (2012) *Equity and quality in education: Supporting disadvantaged students and schools*. Paris: OECD.

Rindermann H (2007) The g-factor of international cognitive ability comparisons: The homogeneity of results in PISA, TIMSS, PIRLS and IQ-tests across nations. *European Journal of Personality* 21(5): 667–706.

Sadler ME (1964) How far can we learn anything of practical value from the study of foreign systems of education? *Comparative Education Review* 7(3): 307–314.

Schütz G, Ursprung HW and Woessmann L (2008) Education policy and equality of opportunity. *Kyklos* 61(2): 279–308.

Van de Werfhorst HG (2014) Changing societies and four tasks of schooling: Challenges for strongly differentiated educational systems. *International Review of Education* 60(1): 123–144.

Van de Werfhorst H and Mijs JJ (2010) Achievement inequality and the institutional structure of educational systems: A comparative perspective. *Annual Review of Sociology* 36: 407–428.

Van Houtte M (2004) Tracking effects on school achievement: A quantitative explanation in terms of the academic culture of school staff. *American Journal of Education* 10(4): 354–388.

Van Houtte M and Stevens PA (2009) Study involvement of academic and vocational students: Does between-school tracking sharpen the difference? *American Educational Research Journal* 46(4): 943–973.

Van Houtte M and Stevens PA (2010) The culture of futility and its impact on study culture in technical/vocational schools in Belgium. *Oxford Review of Education* 36(1): 23–43.

Woessmann L, Luedemann E, Schuetz G and West M (2009) *School Accountability, Autonomy and Choice Around the World*. Cheltenham: Edward Elgar.

Author biographies

Jeroen Lavrijsen is a senior research associate at HIVA–KU Leuven (Research Institute for Work and Society). He investigates the effect of educational system design in the medium-long term (acquisition of qualifications, transition to the labour market) with special attention to patterns of social inequality in these processes.

Ides Nicaise works as a research manager at HIVA–KU Leuven (Research Institute for Work and Society). His research focuses on the economics of education and on poverty and social exclusion.

Appendix 1. Countries participating in the various combinations of assessments under study.

Country	Tracking age	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Argentina	15			X					
Australia	16	X	X				X	X	X
Austria	10	X	X		X	X	X	X	X
Azerbaijan	15						X	X	X
Belgium (Fl.)	12				X	X	X	X	
Belgium (Fr.)	12				X	X			X
Bulgaria	13			X	X	X			X
Canada	16			X					X
Chile	16						X	X	
Chinese Taipei	15	X	X		X	X	X	X	X
Colombia	15	X	X	X					X
Croatia	14,5						X	X	X
Czech Republic	11	X	X	X			X	X	X
Denmark	16	X	X		X	X	X	X	X
Finland	16						X	X	X
France	15			X	X	X			X
Germany	10	X	X	X	X	X	X	X	X
Greece	15			X					
Hong Kong SAR	15	X	X				X	X	
Hungary	11	X	X	X	X	X	X	X	X
Iceland	16			X	X	X			
Indonesia	15				X	X			X
Ireland	15						X	X	X
Israel	15			X	X	X			X
Italy	14	X	X	X	X	X	X	X	X
Japan	15	X	X				X	X	
Korea	14						X	X	
Latvia	16	X	X	X	X	X			
Lithuania	14,5	X	X	X	X	X	X	X	X
Luxembourg	13				X	X			
Netherlands	12	X	X	X	X	X	X	X	X
New Zealand	16	X	X	X	X	X	X	X	X
Norway	16	X	X	X	X	X	X	X	X
Poland	16				X	X	X	X	X
Portugal	15						X	X	X
Qatar	15	X	X		X	X	X	X	X
Romania	14			X	X	X	X	X	X
Russian Federation	14,5	X	X	X	X	X	X	X	X
Singapore	12					X	X	X	X
Slovak Republic	11	X	X	X	X	X	X	X	X
Slovenia	14	X	X	X	X	X	X	X	X
Spain	16				X	X	X	X	X
Sweden	16	X	X	X	X	X	X	X	X
Thailand	15						X	X	
Trinidad and Tobago	11					X			X
Tunisia	16	X	X				X	X	
Turkey	11			X			X	X	
United States	16	X	X			X	X	X	X